Attorney Docket No. YOR920000429US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

Applicant(s): C.C. Aggarwal et al.
Docket No.: YOR920000429US1
Serial No.: 09/686,115
Filing Date: October 11, 2000
Group: 2129
Examiner: Wilbert L. Starks, Jr.

Title: Methods and Apparatus for Outlier Detection for High Dimensional Data Sets

---

APPEAL BRIEF

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313

Sir:

Applicants (hereinafter referred to as "Appellants") hereby appeal the final rejection of claims 1-30 of the above referenced application.

REAL PARTY IN INTEREST

The present application is assigned to International Business Machines Corp., as evidenced by an assignment recorded January 4, 2001 in the U.S. Patent and Trademark Office at Reel 11418, Frame 0019. The assignee, International Business Machines Corp., is the real party in interest.

RELATED APPEALS AND INTERFERENCES

There are no known related appeals and interferences.

STATUS OF CLAIMS

Claims 1-30 are pending in the present application. Claims 1-30 stand rejected under 35 U.S.C. §101. Claims 1-30 are appealed.

STATUS OF AMENDMENTS

An amendment was filed on July 12, 2006, subsequent to the final rejection, but was not entered because it was not deemed to place the application in better form for appeal by materially reducing or simplifying the issues for appeal.

SUMMARY OF INVENTION

The invention relates to outlier detection in high dimensional data and, more particularly, to methods and apparatus for performing such detection in accordance with various high dimensional data domain applications where it is important to be able to find and detect outliers which deviate considerably from the rest of the data (Specification, page 1, lines 4-7).

Claim 1 provides a method of optimizing data mining in a computer. The data mining is performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer. One or more subsets of dimensions and corresponding ranges in the data set which are sparse in density are determined using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions. One or more data points in the data set which contain these subsets of dimensions and corresponding ranges are determined and identified as the one or more outliers in the data set.

By way of example, an illustrative embodiment of the invention of claim 1 is shown in FIG. 3 of the drawings. FIG. 3 shows an overall process for outlier detection according to an embodiment of the present invention. In the set of steps 330-370, an attempt is made to discover those subpatterns of string representations of the database which are very sparsely represented in the database by using a genetic algorithm-based technique. In step 390, the sparsely populated projections for the string encodings together with the corresponding database points are reported (Specification, page 13, line 3 through page 15, line 6).

Claim 10 provides a method of optimizing data mining in a computer. The data mining is performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer. One or more sub-patterns in the data set which have abnormally low presence not due to randomness are identified and mined using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions. One or more records are identified which have the one or more sub-patterns present in them as the one or more outliers.

By way of example, an illustrative embodiment of the invention of claim 10 is shown in FIG. 3 of the drawings. FIG. 3 shows an overall process for outlier detection according to an embodiment of the present invention. In the set of steps 330-370, an attempt is made to discover those subpatterns of string representations of the database which are very sparsely represented in the database by using a genetic algorithm-based technique. In step 390, the sparsely populated projections for the string encodings together with the corresponding database points are reported (Specification, page 13, line 3 through page 15, line 6).

Claim 11 provides an apparatus for optimizing data mining to detect one or more outliers in a high dimensional data set. The apparatus comprises a computer having a memory and a data storage device coupled thereto. The data storage device stores a data store, which has a high dimensional data set of personal attributes. The apparatus further comprises one or more computer programs performed by the computer. The one or more computer programs determine one or more subsets of dimensions and corresponding ranges in the data set which are sparse in density using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions. The one or more computer programs further determine one or more data points in the data set which contain these subsets of dimensions and corresponding ranges, which are identified as the one or more outliers in the data set.

By way of example, an illustrative embodiment of the invention of claim 11 is shown in FIGS. 2 and 3 of the drawings. FIG. 2 shows a hardware implementation suitable for employing outlier detection methodologies according to an embodiment of the present invention. Server 30 contains a large repository of data which is used for the purpose of data mining. The computation is performed by the CPU 32. The data one which the analysis is carried out may already be available

at the server on its disk 36, or it may be specified by the client. Software components including instructions or code for performing the methodologies of the invention may be stored in one or more memory devices and when ready to be utilized, loaded in part or in whole and executed by the CPU (Specification page 12, line 9 through page 13, line 2).

FIG. 3 shows an overall process for outlier detection according to an embodiment of the present invention. In the set of steps 330-370, an attempt is made to discover those subpatterns of string representations of the database which are very sparsely represented in the database by using a genetic algorithm-based technique. In step 390, the sparsely populated projections for the string encodings together with the corresponding database points are reported (Specification, page 13, line 3 through page 15, line 6).

Claim 20 provides an apparatus for optimizing data mining to detect one or more outliers in a high dimensional data set. The apparatus comprises a computer having a memory and a data storage device coupled thereto. The data storage device stores a data store, which has a high dimensional data set of personal attributes. The apparatus further comprises one or more computer programs performed by the computer. The one or more computer programs identify and mine one or more sub-patterns in the data set which have abnormally low presence not due to randomness using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions. The one or more computer programs further identify one or more records which have the one or more sub-patterns present in them as the one or more outliers.

By way of example, an illustrative embodiment of the invention of claim 20 is shown in FIGS. 2 and 3 of the drawings. FIG. 2 shows a hardware implementation suitable for employing outlier detection methodologies according to an embodiment of the present invention. Server 30 contains a large repository of data which is used for the purpose of data mining. The computation is performed by the CPU 32. The data one which the analysis is carried out may already be available at the server on its disk 36, or it may be specified by the client. Software components including instructions or code for performing the methodologies of the invention may be stored in one or more memory devices and when ready to be utilized, loaded in part or in whole and executed by the CPU (Specification page 12, line 9 through page 13, line 2).

FIG. 3 shows an overall process for outlier detection according to an embodiment of the present invention. In the set of steps 330-370, an attempt is made to discover those subpatterns of string representations of the database which are very sparsely represented in the database by using a genetic algorithm-based technique. In step 390, the sparsely populated projections for the string encodings together with the corresponding database points are reported (Specification, page 13, line 3 through page 15, line 6).

Claim 21 provides an article of manufacture comprising a program storage medium readable by a computer and embodying one or more instructions executable by the computer to perform method steps for optimizing data mining. The data mining is performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer. One or more subsets of dimensions and corresponding ranges in the data set which are sparse in density are determined using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions. One or more data points in the data set which contain these subsets of dimensions and corresponding ranges are determined and identified as the one or more outliers in the data set.

By way of example, an illustrative embodiment of the invention of claim 21 is shown in FIGS. 2 and 3 of the drawings. FIG. 2 shows a hardware implementation suitable for employing outlier detection methodologies according to an embodiment of the present invention. Server 30 contains a large repository of data which is used for the purpose of data mining. The computation is performed by the CPU 32. The data one which the analysis is carried out may already be available at the server on its disk 36, or it may be specified by the client. Software components including instructions or code for performing the methodologies of the invention may be stored in one or more memory devices and when ready to be utilized, loaded in part or in whole and executed by the CPU (Specification page 12, line 9 through page 13, line 2).

FIG. 3 shows an overall process for outlier detection according to an embodiment of the present invention. In the set of steps 330-370, an attempt is made to discover those subpatterns of string representations of the database which are very sparsely represented in the database by using a genetic algorithm-based technique. In step 390, the sparsely populated projections for the string

encodings together with the corresponding database points are reported (Specification, page 13, line 3 through page 15, line 6).

Claim 30 provides an article of manufacture comprising a program storage medium readable by a computer and embodying one or more instructions executable by the computer to perform method steps for optimizing data mining. The data mining is performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer. One or more sub-patterns in the data set which have abnormally low presence not due to randomness are identified and mined using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions. One or more records are identified which have the one or more sub-patterns present in them as the one or more outliers.

By way of example, an illustrative embodiment of the invention of claim 11 is shown in FIGS. 2 and 3 of the drawings. FIG. 2 shows a hardware implementation suitable for employing outlier detection methodologies according to an embodiment of the present invention. Server 30 contains a large repository of data which is used for the purpose of data mining. The computation is performed by the CPU 32. The data one which the analysis is carried out may already be available at the server on its disk 36, or it may be specified by the client. Software components including instructions or code for performing the methodologies of the invention may be stored in one or more memory devices and when ready to be utilized, loaded in part or in whole and executed by the CPU (Specification page 12, line 9 through page 13, line 2).

FIG. 3 shows an overall process for outlier detection according to an embodiment of the present invention. In the set of steps 330-370, an attempt is made to discover those subpatterns of string representations of the database which are very sparsely represented in the database by using a genetic algorithm-based technique. In step 390, the sparsely populated projections for the string encodings together with the corresponding database points are reported (Specification, page 13, line 3 through page 15, line 6).

Thus, methods and apparatus of the present invention are provided for outlier detection in databases by determining sparse low dimensional projections, which are used for the purpose of determining which points are outliers. The methodologies of the invention are very relevant in

providing a novel definition of exceptions or outliers for the high dimensional domain of data (Specification, page 17, lines 4-8).

GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

Claims 1-30 stand rejected under 35 U.S.C. §101 as being non-statutory subject matter.

ARGUMENT

Appellants incorporate by reference herein the disclosure of all previous responses filed in the present application, namely, responses dated February 4, 2005, July 26, 2005, and February 15, 2006.

Regarding the §101 rejection of claims 1-30, Appellants respectfully reassert that claims 1-30 are patentable subject matter for at least the reasons presented in Appellants' previous responses, as well as the reasons presented below.

Independent claims 1, 10, 11, 20, 21 and 30 recite patentable subject matter under §101. An algorithm-containing invention is patentable if the invention, as a whole, produces a "useful, concrete, and tangible result," and regardless of the fact that parts of the invention fall within the judicial exceptions of patentable subject matter. *AT&T Corp. v. Excel Comm. Inc.*, 50 U.S.P.Q.2d 1447, 1454 (1999), *State Street Bank & Trust Co. v. Signature Financial Group Inc.*, 49 U.S.P.Q.2d 1596, 1601 (1998), *In re Alappat*, 31 U.S.P.Q.2d 1545, 1557 (1994). Under 35 U.S.C. §101, "[w]hoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title." 35 U.S.C. §101. The judicial exceptions to patentable subject matter include "laws of nature, physical phenomena, and abstract ideas." *Diamond v. Chakrabarty*, 206 U.S.P.Q.2d193, 197 (1980). However, the Federal Circuit has held that an invention is not an unpatentable abstract idea, if the invention produces a "useful, concrete, and tangible result." *In re Alappat*, 31 U.S.P.Q.2d at 1557; *see State Street Bank*, 49 U.S.P.Q.2d at 1601. "The dispositive inquiry is whether the claim *as a whole* is directed to statutory subject matter, it is irrelevant that a claim may contain, as part of the whole, subject matter which would not be patentable by itself." *Alappat*, 31 U.S.P.Q. at 1557.

Examiner contends that because the term "data set" is an abstract idea Appellants' invention is unpatentable. Appellants respectfully disagree with Examiner's application of the general rule of patentable subject matter. First, Appellants assert that the term "data set" is not an abstract idea. An "abstract idea" is an intangible concept or idea, such as, a mathematical algorithm, MPEP §2106(IV)©; a bid, *In re Schrader*, 30 U.S.P.Q.2d 1455, 1458-59; or a bubble hierarchy, *In re Warmerdam*, 31 U.S.P.Q.2d 1754, 1759. Appellants assert that a high dimensional data set of personal attributes stored on a data storage device coupled to a computer is not an abstract idea because the high dimensional data sets are tangible. In a computing sense, the disclosed high dimensional data set of personal attributes can be handled, altered, or destroyed. This is unlike a mathematical formula or a basic law of nature.

Second, even if the high dimensional data sets are considered abstract ideas, the disclosed invention is patentable subject matter because, as in *Alappat*, the invention *as a whole* is directed to statutory subject matter and it is irrelevant that a claim may contain, as part of the whole, unpatentable subject matter. Therefore, optimizing data mining in a computer, the data mining being performed by a computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer, as recited in the independent claims, is patentable, even if it contains abstract concepts such as mathematical equations or other abstract ideas. Since Appellants' invention , as a whole, produces a "useful, concrete, and tangible result," Appellants assert that the claimed invention meets the basic requirements of §101 and is therefore patentable subject matter.

The subject matter of the independent claims is "useful" under 35 U.S.C. §101. An invention is "useful" under §101 if "(I) a person of ordinary skill in the art would immediately appreciate why the invention is useful based on the characteristics of the invention (e.g., properties or applications of a product or process), and (ii) the utility is specific, substantial, and credible." MPEP §2107. Appellants submit that the disclosed invention meets the "useful" requirement because a person of ordinary skill in the art of outlier detection can appreciate the disclosed methods and apparatus for detecting one or more outliers in a data set that is sparse in density. A unique data mining method is provided in the independent claims which applies at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions. The ability to detect

one or more outliers in a high dimensional data set of personal attributes is extremely useful and Appellants assert that the utility of the disclosed invention is specific, substantial, and credible, and therefore patentable.

An invention has "specific utility" if it is specific to the subject matter claimed and can "provide a well-defined and particular benefit to the public." *In re Fisher*, 76 USPQ2d 1225, 1230 (Fed. Cir. 2005). An invention satisfies the "substantial utility" requirement, if the invention's asserted use has a significant and presently available benefit to the public." *Fisher*, 76 U.S.P.Q.2d at 1230. And "credibility" is assessed from "the perspective of one of ordinary skill in the art in view of the disclosure and any other evidence of record that is probative of the applicant's assertions." MPEP §2107(II). An applicant only needs to provide "one credible assertion of specific and substantial utility for each claimed invention to satisfy the utility requirement." *Id.*

The subject matter of the independent claims is specific because it relates to optimizing data mining in a computer. Appellants' disclosure has a well-defined and particular benefit to the public because it involves detecting one or more outliers in a high dimensional data set of personal attributes where one or more subsets of dimensions and corresponding ranges in the data set are sparse in density. Appellants assert that the disclosed invention can detect outliers in data sets where current outlier detection methods may not be able to. For this reason, the subject matter of the independent claims has substantial utility because of its significance to data mining and outlier detection. Furthermore, the subject matter of the independent claims is substantial because the significant benefits of the disclosure are presently available to benefit the public. Finally, the subject matter of the independent claims is credible because a person having ordinary skill in the art of data mining and outlier detection would recognize the specific and substantial utility of Appellants' invention.

The subject matter of the independent claims is "tangible" under 35 U.S.C. §101. An invention is "tangible" under §101 if the "process claim sets forth a practical application of a §101 judicial exception to produce a real-world result." MPEP §2106(IV)©. "A[n] application of a law of nature or mathematical formula to a . . . process may well be deserving of patent protection." *Diamond v. Diehr*, 209 U.S.P.Q. 1, 8. Appellants assert that detecting outliers in a data set which is sparse in density is a tangible result. Although Appellants' utilize mathematical formulas in

detecting one or more outliers in a high dimensional data set of personal attributes, the application of the mathematical formulas produce a real-world result. Examiner contends that Appellants are simply manipulating abstract ideas and that there is no "transformation of real-world data (such as monetary data or heart rhythm data) by some disclosed process." (Office Action pg. 6-7). Appellants respectfully disagree. Detecting one or more outliers in a high dimensional data set, where the one or more subsets of dimensions and corresponding ranges in the data set are sparse in density, is a transformation of real-world data. The transformation is taking an unorganized sparse data set and detecting one or more outliers within the data set. Even if the data set is not specified, there is a tangible result because the outliers in the data set are singled out.

The subject matter of the independent claims is "concrete" under 35 U.S.C. §101. An invention is "concrete" under §101 if the process has "a result that can be substantially repeatable or the process must substantially produce the same result again." MPEP § 2106(IV)(C)(2)(c). The subject matter of the independent claims is substantially repeatable because the disclosed invention will detect the same outliers for a specific high dimensional data set of personal attributes.

Examiner contends that the subject matter of the independent claims is not concrete and tangible because "there are a myriad of possible practical applications." (Office Action pg. 8). Appellants submit that because an invention has a number of possible practical applications does not make it unpatentable. Appellants' invention is specific because it involves optimizing data mining in a computer by detecting one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer. For this reason, Appellants assert that they do not need to specify with over-limiting precision what type of data the disclosed invention is processing.

For at least the reasons given above, independent claims 1, 10, 11, 20, 21, and 30 are patentable because they meet the "useful, concrete, and tangible" requirement of 35 U.S.C. §101. It follows that dependent claims 2-9, 12-19, and 22-29, are also patentable. Appellants respectfully request withdrawal of the §101 rejection of claims 1-30. As such, the application is asserted to be in condition for allowance, and favorable action is respectfully solicited.

Respectfully submitted,

Date: October 23, 2006

Robert W. Griffith
Attorney for Applicant(s)
Reg. No. 48,956
Ryan, Mason & Lewis, LLP
90 Forest Avenue
Locust Valley, NY 11560
(516) 759-4547

CLAIMS APPENDIX

1. A method of optimizing data mining in a computer, the data mining being performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer, the method comprising the steps of:

determining one or more subsets of dimensions and corresponding ranges in the data set which are sparse in density using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions; and

determining one or more data points in the data set which contain these subsets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

2. The method of claim 1, wherein a range is defined as a set of contiguous values on a given dimension.

3. The method of claim 1, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

4. The method of claim 3, wherein the sparsity coefficient measure $S(D)$ is defined as $\frac{n(D) - N * f^k}{\sqrt{N * f^k * (1 - f^k)}}$, where k represents the number of dimensions in the data set, f represents the fraction of data points in each range, N is the total number of data points in the data set, and $n(D)$ is the number of data points in a set of dimensions D.

5. The method of claim 3, wherein a given sparsity coefficient measure is inversely proportional to the number of data points in a given set of dimensions and corresponding ranges.

6. The method of claim 1, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

7. The method of claim 6, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

8. The method of claim 6, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

9. The method of claim 6, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

10. A method of optimizing data mining in a computer, the data mining being performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer, the method comprising the steps of:

identifying and mining one or more sub-patterns in the data set which have abnormally low presence not due to randomness using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions; and

identifying one or more records which have the one or more sub-patterns present in them as the one or more outliers.

11. Apparatus for optimizing data mining to detect one or more outliers in a high dimensional data set comprising:

a computer having a memory and a data storage device coupled thereto, wherein the data storage device stores a data store, the data store having a high dimensional data set of personal attributes; and

one or more computer programs, performed by the computer, for: (I) determining one or more subsets of dimensions and corresponding ranges in the data set which are sparse in density

using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions; and (ii) determining one or more data points in the data set which contain these subsets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

12. The apparatus of claim 11, wherein a range is defined as a set of contiguous values on a given dimension.

13. The apparatus of claim 11, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

14. The apparatus of claim 13, wherein the sparsity coefficient measure $S(D)$ is defined as $\frac{n(D) - N * f^k}{\sqrt{N * f^k * (1 - f^k)}}$, where k represents the number of dimensions in the data set, f represents the fraction of data points in each range, N is the total number of data points in the data set, and $n(D)$ is the number of data points in a set of dimensions D.

15. The apparatus of claim 13, wherein a given sparsity coefficient measure is inversely proportional to the number of data points in a given set of dimensions and corresponding ranges.

16. The apparatus of claim 11, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

17. The apparatus of claim 16, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

18. The apparatus of claim 16, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

19. The apparatus of claim 16, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

20. Apparatus for optimizing data mining to detect one or more outliers in a high dimensional data set comprising:

a computer having a memory and a data storage device coupled thereto, wherein the data storage device stores a data store, the data store having a high dimensional data set of personal attributes; and

one or more computer programs, performed by the computer for: (I) identifying and mining one or more sub-patterns in the data set which have abnormally low presence not due to randomness using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions; and (ii) identifying one or more records which have the one or more sub-patterns present in them as the one or more outliers.

21. An article of manufacture comprising a program storage medium readable by a computer and embodying one or more instructions executable by the computer to perform method steps for optimizing data mining, the data mining being performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer, the method comprising the steps of:

determining one or more subsets of dimensions and corresponding ranges in the data set which are sparse in density using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions; and

determining one or more data points in the data set which contain these subsets of dimensions and corresponding ranges, the one or more data points being identified as the one or more outliers in the data set.

22. The article of claim 21, wherein a range is defined as a set of contiguous values on a given dimension.

23. The article of claim 21, wherein the sets of dimensions and corresponding ranges in which the data is sparse in density is quantified by a sparsity coefficient measure.

24. The article of claim 23, wherein the sparsity coefficient measure $S(D)$ is defined as $\frac{n(D) - N * f^k}{\sqrt{N * f^k * (1 - f^k)}}$, where k represents the number of dimensions in the data set, f represents the fraction of data points in each range, N is the total number of data points in the data set, and $n(D)$ is the number of data points in a set of dimensions D.

25. The article of claim 23, wherein a given sparsity coefficient measure is inversely proportional to the number of data points in a given set of dimensions and corresponding ranges.

26. The article of claim 21, wherein a set of dimensions is determined using an algorithm which uses the processes of solution recombination, selection and mutation over a population of multiple solutions.

27. The article of claim 26, wherein the process of solution recombination comprises combining characteristics of two solutions in order to create two new solutions.

28. The article of claim 26, wherein the process of mutation comprises changing a particular characteristic of a solution in order to result in a new solution.

29. The article of claim 26, wherein the process of selection comprises biasing the population in order to favor solutions which are more optimum.

30. An article of manufacture comprising a program storage medium readable by a computer and embodying one or more instructions executable by the computer to perform method steps for optimizing data mining, the data mining being performed by the computer to detect one or more outliers in a high dimensional data set of personal attributes stored on a data storage device coupled to the computer, the method comprising the steps of:

identifying and mining one or more sub-patterns in the data set which have abnormally low presence not due to randomness using an algorithm capable of utilizing at least one of the processes of solution recombination, selection and mutation over a population of multiple solutions; and

identifying one or more records which have the one or more sub-patterns present in them as the one or more outliers.

EVIDENCE APPENDIX

None.

RELATED PROCEEDINGS APPENDIX

None.